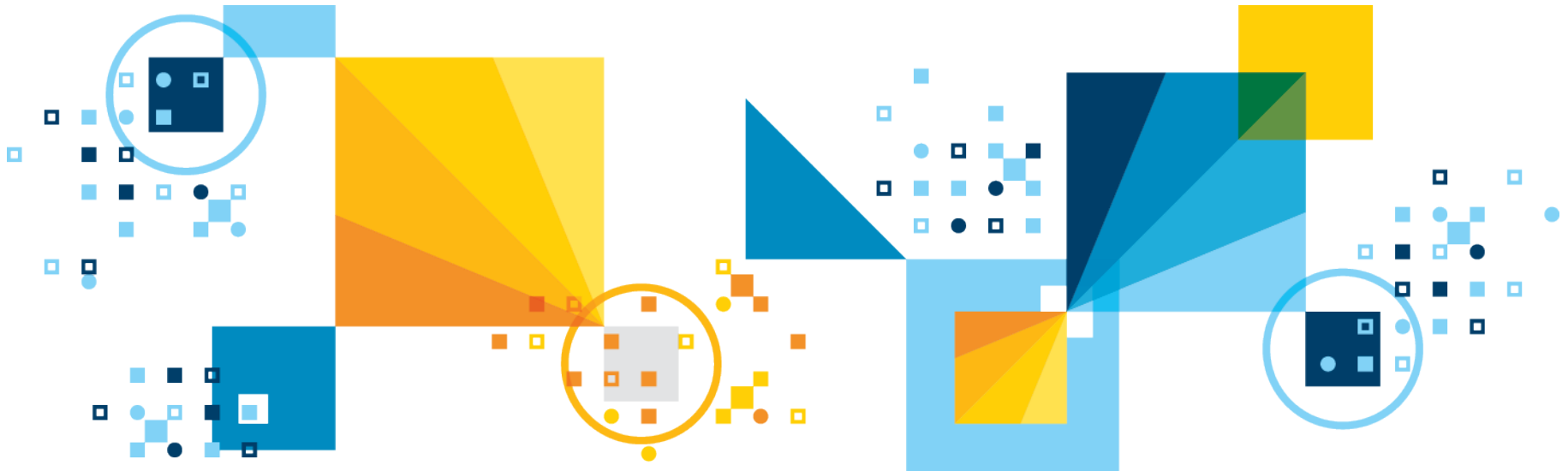


Dec 15<sup>th</sup>, 2016

# Welcome to the IBM IIS Tech Talk

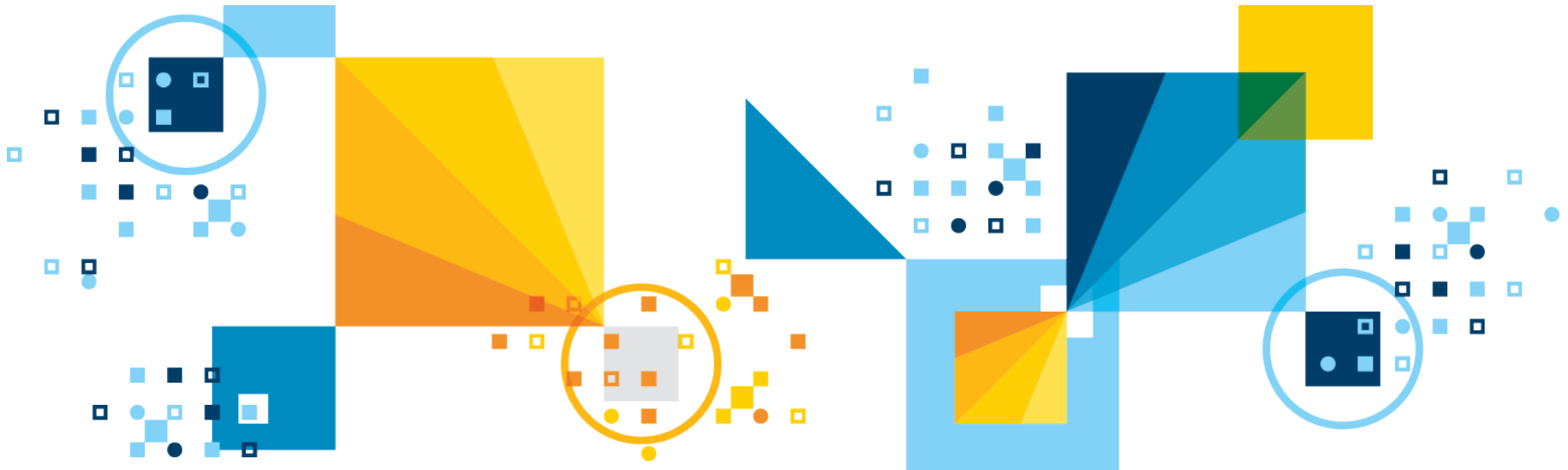
## Data Quality in Information Analyzer



Dec 15<sup>th</sup>, 2016

# Information Analyzer Data Quality Deep Dive

Yannick Saillet – Software Architect



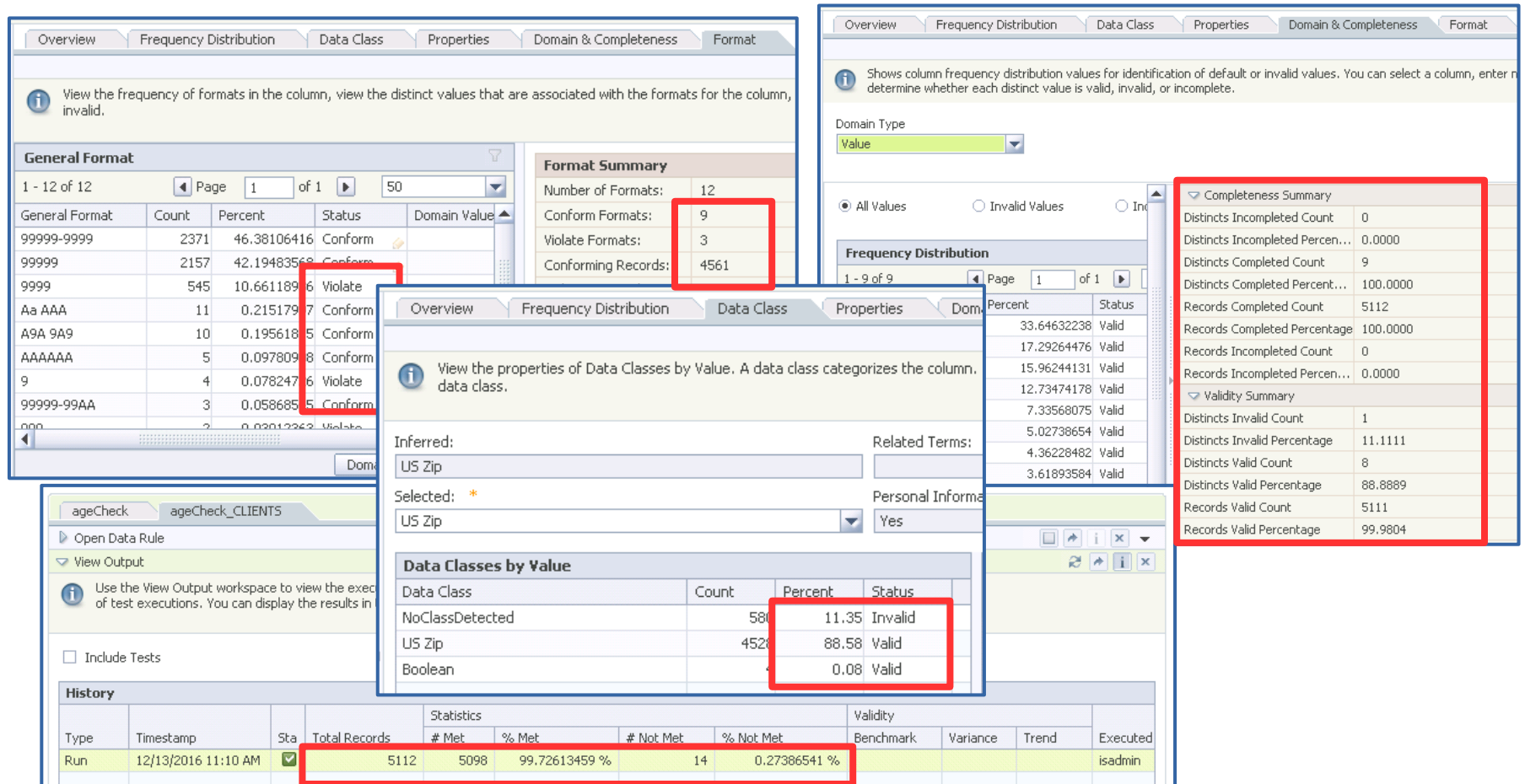
# AGENDA

- Data Quality in Information Analyzer in the past
- Data Quality in IA 11.5 thin client
- The Data Quality Framework
- Out of the box data quality problems detection
- Reporting
- Q&A

# Analytics Platform Services - Disclaimer

- IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion.
- Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.
- The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.
- The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.
- Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

# Data Quality in Information Analyzer in the past



The screenshot displays several overlapping windows from the IBM Information Analyzer interface, illustrating data quality information spread across multiple screens. Red boxes highlight specific data points:

- Format Summary:**

Number of Formats:	12
Conform Formats:	9
Violate Formats:	3
Conforming Records:	4561
- Format Table:**

General Format	Count	Percent	Status	Domain Value
99999-9999	2371	46.38106416	Conform	
99999	2157	42.19483568	Conform	
9999	545	10.661189	Violate	
Aa AAA	11	0.215179	Conform	
A9A 9A9	10	0.195618	Conform	
AAAAAA	5	0.097809	Conform	
9	4	0.078247	Violate	
99999-99AA	3	0.058685	Conform	
000	2	0.03912262	Violate	
- Completeness Summary:**

Distincts Incompleted Count	0
Distincts Incompleted Percent...	0.0000
Distincts Completed Count	9
Distincts Completed Percent...	100.0000
Records Completed Count	5112
Records Completed Percentage	100.0000
Records Incompleted Count	0
Records Incompleted Percent...	0.0000
- Validity Summary:**

Distincts Invalid Count	1
Distincts Invalid Percentage	11.1111
Distincts Valid Count	8
Distincts Valid Percentage	88.8889
Records Valid Count	5111
Records Valid Percentage	99.9804
- Data Classes by Value:**

Data Class	Count	Percent	Status
NoClassDetected	58	11.35	Invalid
US Zip	452	88.58	Valid
Boolean		0.08	Valid
- History Table:**

Type	Timestamp	Sta	Total Records	# Met	% Met	# Not Met	% Not Met	Validity	Benchmark	Variance	Trend	Executed
Run	12/13/2016 11:10 AM	✓	5112	5098	99.72613459 %	14	0.27386541 %					isadmin

- Many data quality related information widespread on many different screens: (Domain validity, formats, data classes, data rules, key referential integrity, ...)

# The challenge of measuring Data Quality

BANK1.BANK_CLIENTS						
CLIENT_ID	NAME [CHAR(128 OCT)]	ADDRESS [CHAR(128 OCTETS)]	ZIP [CHAR(10 O)]	AGE [RI]	GENDE	
1578	Tyler O Abatemarco	Witton Birmingham	77019-6813	85.0	M	
1579	Debra M Pedrick	Kings Heath Birmingham	80110-4998	78.0		
1580	Cassandra R Ker	Kings Heath Birmingham	33781-2153	38.0		
1581	Johnson Y Foltz	351-353 Lea Bridge Road Leyt	34787	75.0		
1581	Johnson Y Foltz	351-353 Lea Bridge Road Leyt	34787	75.0		
1582	Mary H Jacques	121 W Canal Stre Unit 53 Leed	00000-0000	62.0		
1583	Pual G Fowler	36 GRAVELLY INDUSTRIAL PAR	19175-5490	59.0		
1584	Thoang D Meyers	Woodilee Road Kirkintilloch Gl	78723-6199	86.0		
1584	Thoang D Meyers	Woodilee Road Kirkintilloch Gl	78723-6199	86.0		
1585	Janet M Alcazar	CLITTAFORD RD SOUTHWAY	34741-5027	24.0	M	
1586	Richard A Pringle	257 Great Lister Street P O BOX	32114-3851	57.0	F	
1587	Christina P Gee	EASTLEIGH HAMPSHIRE S050	86047-2538			
1587	Christina P Gee	EASTLEIGH HAMPSHIRE S050	86047-2538			
1588	Maurits Q Schuller	2560 MCMULLEN BOOTH RD	33761-4100			
1589	Lillian R Isaac	Fleetesex Hampshire	33309-3421			
1590	Lucy Y Adler	Blucher Street Birmingham	27406-6355			
1590	Lucy Y Adler	Blucher Street Birmingham	27406-6355			
1591	Cory J Gardner	5 Oxford Road Newbury	24742-0460			

% duplicate rows = 2%  
% Referential integrity violation = 3%

Data rule #1: 15% failed  
Data rule #2: 5% failed  
Data rule #3: 1% failed

% unique=73%  
% null=5%  
Nb formats=3

% unique=99%  
% null=2%  
% data class violations=7%



BANK2.BANK_CUSTOMERS					
CUSTOMER_ID	NAME [CHAR(128 O)]	ADDRESS [CHAR(128 OCTETS)]	ZIP [CHAR(10 O)]	CREDIT_RATING	
1320	Herman C Trappe	GASGOIGNE ROAD ESSEX	28269-7613	690	
1321	Ada O Larose	Borehamwood P O BOX 150 Her	34474	690	
1322	Hermine V Zoellner	Grafton Road West Bromwich	29033	616	
1323	Wess T Amado	4 New Square Feltham	32805	616	
1324	Gilmer Y Beach	ABERCRAVE CAERBONT SWAF	27603-1421	616	
1325	Vomni K Littlepage	FARNCOMBE ROAD WEST SUS	34209	816	
1326	Cesarea R Thiel	Long Eaton Nottingham	30268-2413	816	
1327	Phil X Trevino	60 Frederick Street Birmingham	32085-1048	816	
1328	Hermine A Tews	502 HONEYPOT LANE STANMC	77229-4071	816	
1329	Graig Q Pearson	502 HONEYPOT LANE STANMC	32254	775	
1330	Jameson T Watson	Witton Birmingham	30336-2435	775	
1331	Yvan F Decamp	HOLFORD DRIVE HOLFORD BIR	77080-2736	775	
1332	Eron B Matsukura	35 Livery Street Birmingham	34773-9109	775	
1333	Sueann S Paranjpe	Hall Weston CAMBS	85304	644	
1334	Lewis S Leland	Gulldhill Lane Leicester	85259-7008	644	
1335	Lorvel I Galligan	Canal Road Leeds	54235-0736	644	
1336	Freddit I Berge	Sherbourne Drive Tilbrook Milto	78028-5199	644	
1337	Antony I Lehrkind	Oldfield Road Hampton	70810	673	
1338	Roselle I Gagnon	Bordesley Birmingham	33312-3109	673	

% duplicate rows = 4%  
% Referential integrity violation = 1%

Data rule #4: 20% failed  
Data rule #5: 6% failed

% unique=87%  
% null=15%  
Nb formats=5

% unique=93%  
% null=5%  
% data class violations=2%

- Which data set has the better quality?
- What are the quality problems I need to resolve first to increase the quality?
- => How to consolidate heterogeneous quality metric into a single metric?

# Data Quality in Information Analyzer 11.5 Thin Client

The screenshot displays the IBM InfoSphere Information Analyzer 11.5 Thin Client interface. The top navigation bar includes 'Workspaces', 'Data catalog', and 'Connections'. The main workspace shows a data set named '360 degrees view customers - BAN...' with a 'DATA QUALITY' score of 98% highlighted in a red box. The interface also displays a list of data quality violations, including 'DATA CLASS VIOLATIONS' (1335 findings | 1%), 'SUSPECT VALUES' (346 findings | 0%), 'RULE VIOLATIONS: DEMOGRAPHIC\_JOINEDACCOUNT\_DEFINITION' (318 findings | 0%), and 'INCONSISTENT USAGE OF UPPER AND LOWER CASES.' Each violation has an 'Ignore' toggle and a dropdown menu.

Category	Findings	Percentage
DATA CLASS VIOLATIONS	1335	1%
SUSPECT VALUES	346	0%
RULE VIOLATIONS: DEMOGRAPHIC_JOINEDACCOUNT_DEFINITION	318	0%
INCONSISTENT USAGE OF UPPER AND LOWER CASES.		

- A **consolidated** single **data quality score** is computed for each data set and field.
- It makes it easy to compare the quality of different data sets or data fields.
- The full data quality details explaining the scores are still captured.

## How it works: Data Quality Framework

- Extensible<sup>(\*)</sup> framework where different algorithms specialized in detecting specific types of data quality problems can be plugged.
- The framework will run the algorithms on the data and consolidate the findings into:
  - A normalized data quality score for the whole data set, each data field, or even each row and value.
  - A distribution of the frequency of each type of problem is found for the whole data set or each individual data field.
- The data quality scores are computed in a standardized way
  - No matter how many algorithms are used.
  - No matter what type of problems these algorithms search.
  - No matter how the algorithms find those problems.

(\*): As of the current version, the extension point to integrate new algorithms is not yet public. A public extension point may be considered in the future.



## How it works: Data Quality Framework (cont)

- Each algorithm is a so called “Quality Scanner” which can search for the existence of one or multiple types of data quality problems within a single cell, a complete row, a complete column or the data set as a whole.
- The scanner declares to the framework:
  - The type of problems it searches,
  - Whether it search for problems within individual cells, rows, columns or the data set as a whole.
  - How many passes over the data it requires.
- The framework invokes the scanners.
- When a scanner detects a data quality problem, it annotates the cell, row, column or data set where it found the problem with:
  - The type of problem which has been identified.
  - The confidence that the problem is not a false positive.
- The framework consolidates the annotations produced by each scanner into consistent data quality scores.

# Computing Data Quality Scores

- Data quality score
  - **Estimate of the proportion of reliable data values in the given dataset**
  - It is represented on the percentage scale
  
- Cell score: estimate proportion of all data quality problems per cell
  - $q$ : data quality problem that is applicable for a given cell
  - $prev(q)$ : prevalence for the data quality problem identified by  $q$  (prevalence=estimation of the % of values having that problem)
  - $conf(q)$ : confidence for the data quality problem identified by  $q$  (confidence=probability that the detected problem is not a false positive)
  - $Q_{cell}$ : the number of all data quality problems that are applicable for the given cell
  - $Score(cell)$ : data quality score computed for the given cell

$$Score(cell) = \prod_{q=0}^{Q_{cell}} (1 - prev(q) * conf(q))$$

# Column and Total Data Quality Scores

- Column score: estimate the column proportion of reliable data values

- *col*: data set column of interest
- *row*: data set row numbered from 1 to *N*
- *Score (col)*: data quality score computed for the given column

$$Score(col) = \frac{1}{N} \sum_{row=1}^N Score(row, col)$$

- Dataset score: estimate of the total proportion of reliable data values

- *M*: the number of columns in the given data-set
- *weight(col)*: pre-specified weight for each column with the total *W*
- *Score (data-set)*: data quality score computed for the whole data-set

$$Score(dataset) = \frac{1}{W} \sum_{col=1}^M Score(col) * weight(col)$$

# Example

Cust ID	Name	Age	Phone	Gender
62413	Lucy V Adler	32	334-555-6633	F
62414	Cory J Gardner	25	903-222-1255	F
62414	Mary H Jacques	18	777-156-9836	F
62415	Pdsaojfsadpoifj	46	xxxx	M
62416	Shaun Q Dunda	156	904-555-2940	M
62417	Carol T Schwartz	22	804-555-3164	F
62418	HARRIS LAURENT	36	785-555-5835	-

# Example

Cust ID	Name	Age	Phone	Gender
62413	Lucy V Adler	32	334-555-6633	F
<b>62414</b>	Cory J Gardner	25	903-222-1255	F
<b>62414</b>	Mary H Jacques	18	777-156-9836	F
62415	<b>Pdsaojfsadpoifj</b>	46	<b>xxxx</b>	M
62416	Shaun Q Dunda	<b>156</b>	904-555-2940	M
62417	Carol T Schwartz	22	804-555-3164	F
62418	<b>HARRIS LAURENT</b>	36	785-555-5835	-

# Example

Cust ID	Name	Age	Phone	Gender
62413	V Adler	37	55-6633	F
<b>62414</b>	Gardner		2-1255	F
<b>62414</b>	H Jacqu		6-9836	F
62414	<b>Pdsaojfsadpoij,</b>		<b>xxxx</b>	M
62414	Shaun Q Dunda	<b>156</b>	904	M
62414	Carol T Schwartz	22	804	F
62418	<b>HARRIS LAURENT</b>	36	785-5	F

Uniqueness violation  
Conf:100%

Suspect value  
Conf:90%

Outlier  
Conf:95%

Data Class Violation  
Conf:100%

Inconsistent Case  
Conf:98%

Missing Value  
Conf:100%

# Example

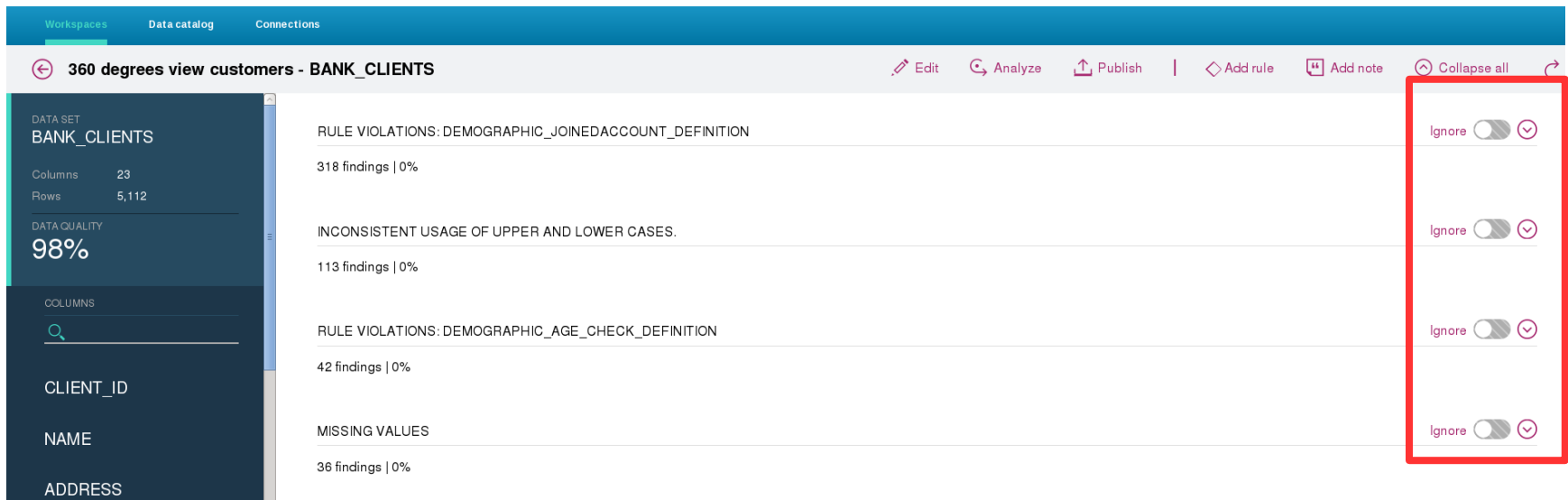
Score: 71%      Score: 73%      Score: 86%      Score: 85%      Score: 85%

Cust ID	Name	Age	Phone	Gender
62413	Lucy V Adler	32	334-555-6633	F
<b>62414</b>	Cory J Gardner	25	903-222-1255	F
<b>62414</b>	Mary H Jacques	18	777-156-9836	F
62415	<b>Pdsaojfsadpoifj</b>	46	<b>xxxx</b>	M
62416	Shaun Q Dunda	<b>156</b>	904-555-2940	M
62417	Carol T Schwartz	22	804-555-3164	F
62418	<b>HARRIS LAURENT</b>	36	785-555-5835	-

**Data Set**  
**Score: 80%**

# Data quality problems detected out of the box

- Information Analyzer includes a collection of out of the box algorithms specialized in detecting different types of data quality problems.
- You can choose to ignore some of these data quality problems in the UI.
  - Problems can be ignored for a single data field only, or for the whole data set.
  - Ignored problems won't impact the score and are not going to be searched in consequent analysis.



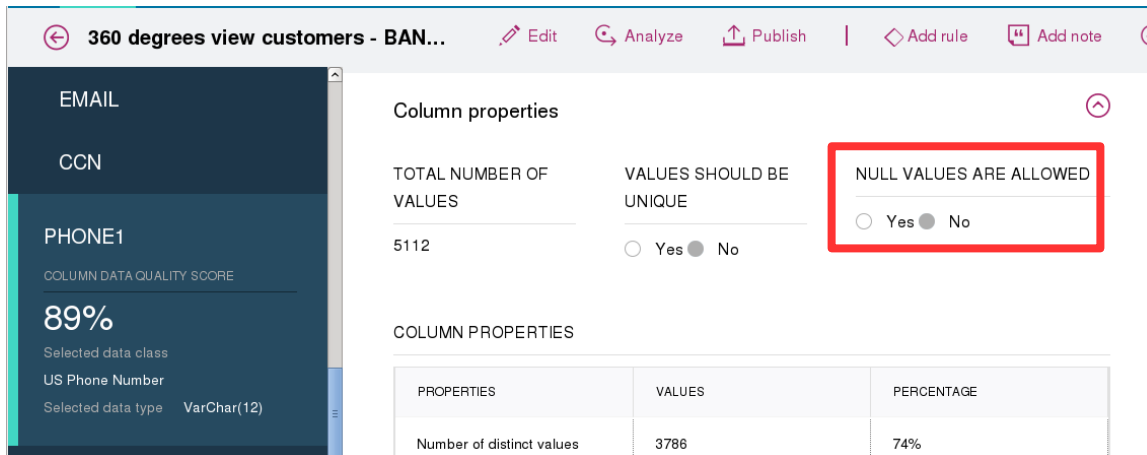
The screenshot displays the IBM Analytics interface for a data set named 'BANK\_CLIENTS'. The left sidebar shows the data set name, 23 columns, 5,112 rows, and a data quality score of 98%. The main panel lists several rule violations, each with a corresponding 'Ignore' toggle and a dropdown arrow. A red box highlights these 'Ignore' controls for four specific rule violations:

- RULE VIOLATIONS: DEMOGRAPHIC\_JOINEDACCOUNT\_DEFINITION (318 findings | 0%)
- INCONSISTENT USAGE OF UPPER AND LOWER CASES. (113 findings | 0%)
- RULE VIOLATIONS: DEMOGRAPHIC\_AGE\_CHECK\_DEFINITION (42 findings | 0%)
- MISSING VALUES (36 findings | 0%)



# Data quality problems detected out of the box: Missing Values

- Checks for missing values in data fields where missing values are not supposed to be expected.
- If a data field is declared by the user as non nullable, all null and empty values will be reported as a quality problem with a confidence of 100%.
- The nullability flag of the data fields can be controlled in the UI:



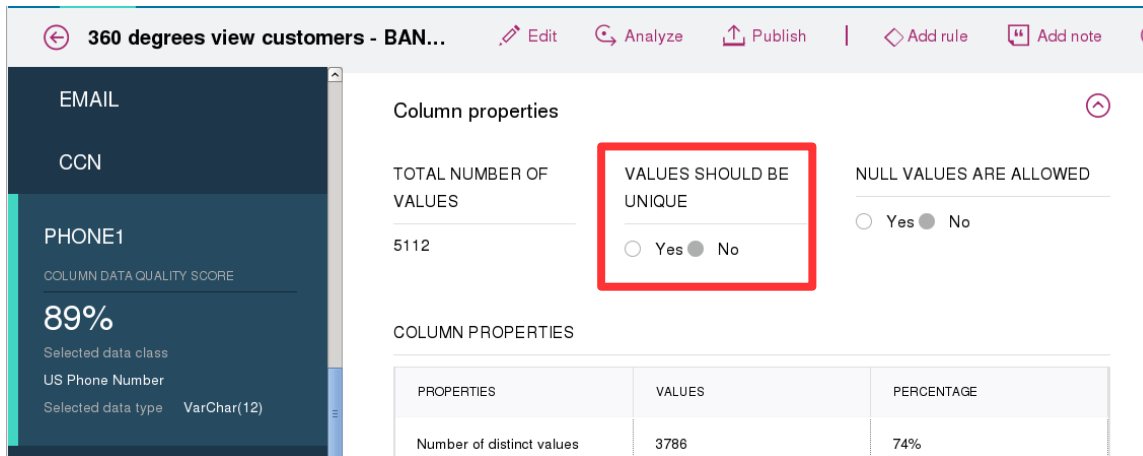
The screenshot shows the '360 degrees view customers - BAN...' interface. On the left, a sidebar displays the column 'PHONE1' with a 'COLUMN DATA QUALITY SCORE' of 89%. The main panel shows 'Column properties' for 'PHONE1' with a total of 5112 values. A red box highlights the 'NULL VALUES ARE ALLOWED' setting, which is currently set to 'No'. Below this, a table shows 'COLUMN PROPERTIES' with 3786 distinct values representing 74% of the total.

PROPERTIES	VALUES	PERCENTAGE
Number of distinct values	3786	74%

- If no information is available about the nullability of a data field, null and empty values are only reported as a quality problem if the majority of the values in the data field are not missing. The confidence of the reported problem will be proportional to the percentage of non missing values in the data field.

# Data quality problems detected out of the box: Uniqueness Violations

- Checks for duplicate values in data fields supposed to contain only unique values.
- If a data field is declared by the user as unique, all non unique values will be reported as a quality problem with a confidence of 100%.
- The uniqueness flag of the data fields can be controlled in the UI:



The screenshot shows the 'Column properties' configuration for the 'PHONE1' column. The 'VALUES SHOULD BE UNIQUE' checkbox is checked, and the 'NULL VALUES ARE ALLOWED' radio button is set to 'No'. The 'TOTAL NUMBER OF VALUES' is 5112. The 'COLUMN DATA QUALITY SCORE' is 89%.

PROPERTIES	VALUES	PERCENTAGE
Number of distinct values	3786	74%

- If no information is available about the uniqueness of a data field, duplicate values are only reported as a quality problem if the majority of the values in the data field are unique. The confidence of the reported problem will be proportional to the percentage of unique values in the data field.

# Data quality problems detected out of the box: Invalid formats

- Checks for values having a format marked by the user as invalid for its data field.
- The validity of the formats can be controlled in the UI:

360 degrees view customers - BANK\_CLIENTS

[Edit](#) | [Analyze](#) | [Publish](#) | [Add rule](#) | [Add note](#) | [Collapse all](#)

NAME

ADDRESS

ZIP

COLUMN DATA QUALITY SCORE

88%

Selected data class US Zip

Selected data type VarChar(10)

AGE

GENDER

MARITAL\_STATUS

PROFESSION

NBR\_YEARS\_CLI

SAVING\_ACCOUNT

### Formats

**FOUND FORMATS**

12 findings of 5112 records

---

**VALID FORMATS**

12 formats | 5112 records | 100%

**MOST COMMON FORMAT**

99999-9999

---

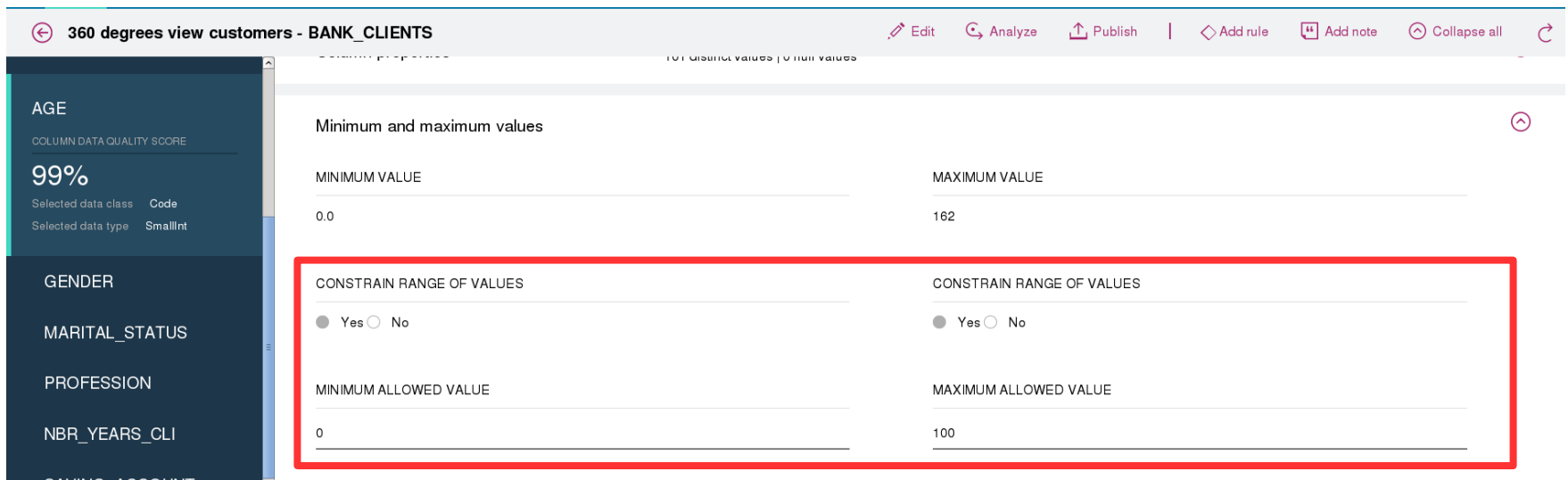
**INVALID FORMATS**

0 formats | 0 records | 0%

FORMAT	COUNT	PERCENTAGE	VALID
99999-9999	2371	46.38%	<input checked="" type="checkbox"/>
99999	2157	42.19%	<input checked="" type="checkbox"/>
9999	545	10.66%	<input checked="" type="checkbox"/>
Aa AAA	11	0.22%	<input checked="" type="checkbox"/>
A9A 9A9	10	0.20%	<input checked="" type="checkbox"/>

# Data quality problems detected out of the box: Minimum / Maximum Range Violations

- Checks for values being outside the valid minimum/maximum range defined by the user for the data field.
- The validity of the formats can be controlled in the UI:



The screenshot displays the IBM Analytics interface for the '360 degrees view customers - BANK\_CLIENTS' dataset. On the left, a sidebar shows the 'AGE' field with a 99% data quality score and a 'SmallInt' data type. The main panel shows the 'Minimum and maximum values' configuration for the AGE field. The current minimum value is 0.0 and the maximum value is 162. Below this, there are two 'CONSTRRAIN RANGE OF VALUES' sections, each with a 'Yes' radio button selected. The first section has a minimum allowed value of 0 and a maximum allowed value of 100. The second section has a minimum allowed value of 0 and a maximum allowed value of 100. A red box highlights these two sections.

Minimum and maximum values	
MINIMUM VALUE	MAXIMUM VALUE
0.0	162

CONSTRRAIN RANGE OF VALUES	
<input checked="" type="radio"/> Yes <input type="radio"/> No	<input checked="" type="radio"/> Yes <input type="radio"/> No
MINIMUM ALLOWED VALUE	MAXIMUM ALLOWED VALUE
0	100

CONSTRRAIN RANGE OF VALUES	
<input checked="" type="radio"/> Yes <input type="radio"/> No	<input checked="" type="radio"/> Yes <input type="radio"/> No
MINIMUM ALLOWED VALUE	MAXIMUM ALLOWED VALUE
0	100

# Data quality problems detected out of the box: Invalid Data Types

- Checks for values incompatible with the inferred/selected data type of their data field.
- The data type of the data fields can be specified in the UI:

IBM InfoSphere Information Analyzer

Administrator IIS | About | Help | Sign out | IBM

Workspaces | Data catalog | Connections

← 360 degrees view customers - BANK\_CLIENTS | Edit | Analyze | Publish | Add rule | Add note | Collapse all | Refresh

**AGE**  
COLUMN DATA QUALITY SCORE  
**99%**  
Selected data class: Code  
Selected data type: SmallInt

**GENDER**

**MARITAL\_STATUS**

**PROFESSION**

**NBR\_YEARS\_CLI**

**SAVING\_ACCOUNT**

**ONLINE\_ACCESS**

Data types

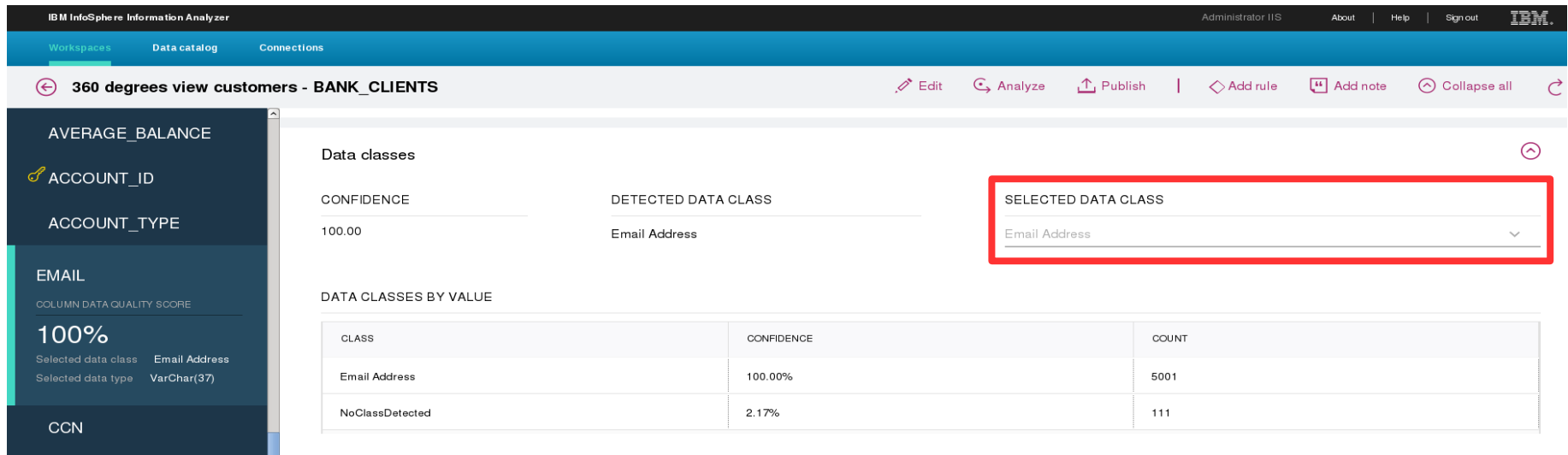
DATA TYPE AS DEFINED ON THE SOURCE		BEST DATA TYPE MATCHING THE VALUES	
VARCHAR(1024)		SMALLINT	
MINIMUM LENGTH	MAXIMUM LENGTH	AVERAGE LENGTH	
3	5	3.98	

SELECTED DATA TYPE  
SMALLINT

TYPE	COUNT	PERCENTAGE
SMALLINT	1	0.02%
TINYINT	5111	99.98%

# Data quality problems detected out of the box: Data Class Violations

- Checks for values which don't match the data class selected for their data field.
- The selected data class of the data fields can be specified in the UI:



IBM InfoSphere Information Analyzer

Administrator IIS | About | Help | Sign out

Workspaces | Data catalog | Connections

← 360 degrees view customers - BANK\_CLIENTS | Edit | Analyze | Publish | Add rule | Add note | Collapse all | Refresh

**AVERAGE\_BALANCE**

ACCOUNT\_ID

ACCOUNT\_TYPE

**EMAIL**

COLUMN DATA QUALITY SCORE

**100%**

Selected data class: Email Address  
Selected data type: VarChar(37)

CCN

Data classes

CONFIDENCE	DETECTED DATA CLASS	SELECTED DATA CLASS
100.00	Email Address	Email Address

DATA CLASSES BY VALUE

CLASS	CONFIDENCE	COUNT
Email Address	100.00%	5001
NoClassDetected	2.17%	111

## Data quality problems detected out of the box: Inconsistent use of upper and lower case

- Checks for values in a data field which have a different use of case (All upper case, vs. all lower case, vs. Name case, vs. Sentence case)
- Information Analyzer will recognize if the majority of the values of a data field use a consistent case usage, and mark values using a different case as a potential data quality problem.
- Ex:

Lucy V Adler

Cory J Gardner

Mary H Jacques

Shaun Q Dunda

Carol T Schwartz

**HARRIS LAURENT**

## Data quality problems detected out of the box: Inconsistent representation of missing values

- Checks for data fields containing both NULL and empty of spaces values, as they indicate different and non standardized representations of the missing values.
- Ex:

NULL  
NULL  
NULL  
**<EMPTY>**  
**<SPACES>**  
NULL  
NULL  
NULL



# Data quality problems detected out of the box:

## Suspect Values

- Checks for values which are either numerical or categorical outliers, or values which seem not to be of the same domain as the other values of the same data field.
- For string values, the system will use advanced algorithms to detect patterns in the properties of the values, such as value length, % of letter/digits, nb of words, recurrent words, recurrent formats, used characters, etc... in order to detect values that look different than the majority of the values of the same data field
- Ex:

Name	Age
Lucy V Adler	32
Cory J Gardner	25
Mary H Jacques	18
<b>Pdsaojfsadpoifj</b>	46
Shaun Q Dunda	<b>156</b>
Carol T Schwartz	22

## Data quality problems detected out of the box: Violations of Correlation

- Checks for values which are plausible in the context of their data field, but not in combination of the other values of the same row.
- Information Analyzer will identify numerical and categorical correlations between columns and mark the values that violate those correlations as potential data quality problems.
- Note: as this algorithm is more expensive in term of computation time, it is not enabled by default and needs to be enabled first:

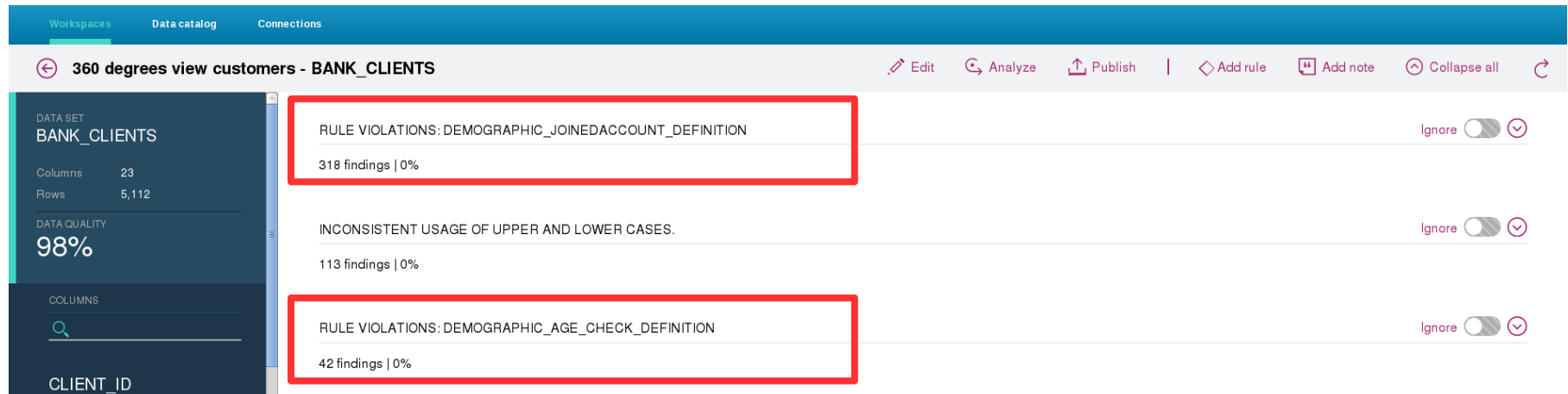
```
ASBServer\bin\iisAdmin.bat -set -key com.ibm.iis.ia.server.dqa.includeCorrelations -value true
```

- Ex:

City	State
San Francisco	CA
Las Vegas	NV
San Francisco	CA
<b>San Francisco</b>	<b>NV</b>
Las Vegas	NV
San Francisco	CA

# Data quality problems detected out of the box: Data Rule Violations

- If data rules or quality rules are defined on the analyzed data set, their logic is going to be automatically used during the data quality analysis if the rule fulfills the following conditions:
  - All its variables are only bound to data fields of the same data set or literals.
  - Its logic doesn't contain any lookup or aggregation.
- Each used data rule definition is reported as a separate data quality problem type.

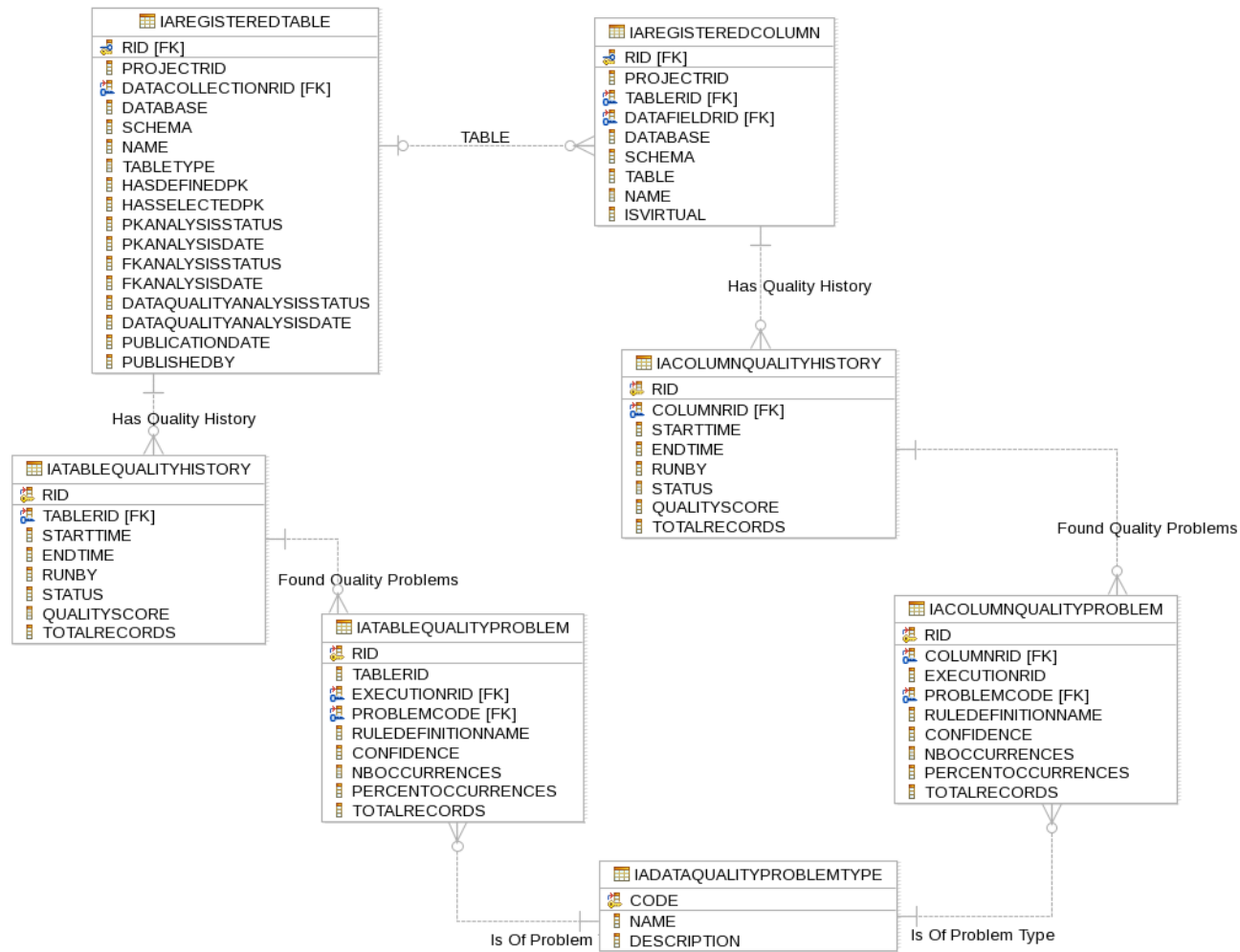


The screenshot shows the IBM Analytics interface for the 'BANK\_CLIENTS' dataset. The left sidebar displays the dataset name, column count (23), row count (5,112), and a data quality score of 98%. The main area lists three data quality issues, each with a red box highlighting the rule name and finding count:

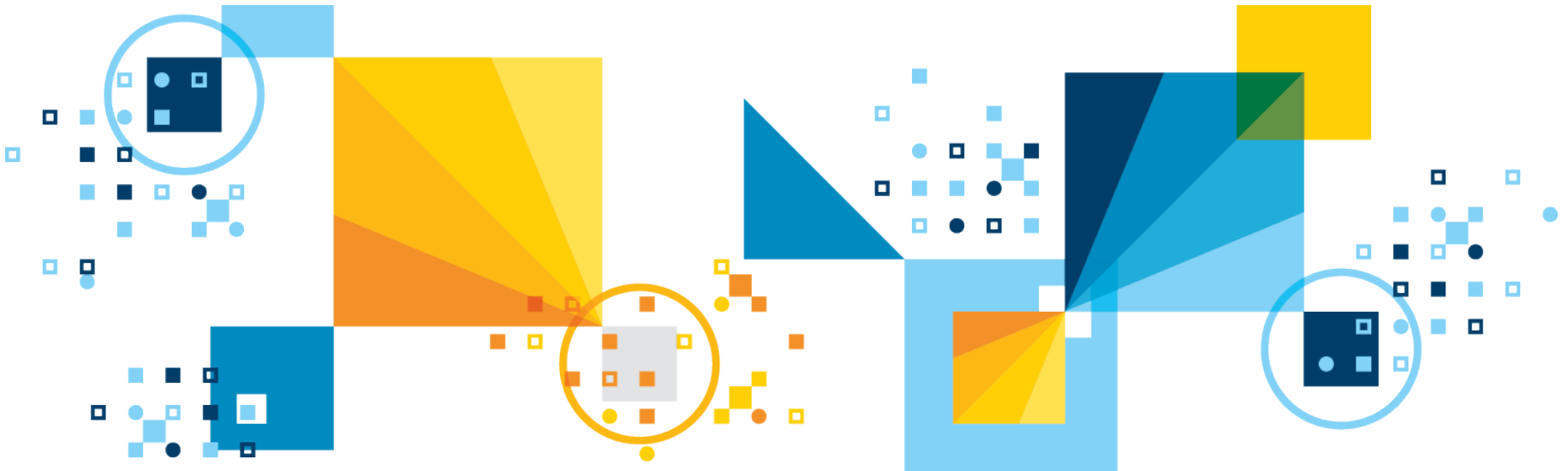
Rule Name	Findings	0%
RULE VIOLATIONS: DEMOGRAPHIC_JOINEDACCOUNT_DEFINITION	318 findings	0%
INCONSISTENT USAGE OF UPPER AND LOWER CASES.	113 findings	0%
RULE VIOLATIONS: DEMOGRAPHIC_AGE_CHECK_DEFINITION	42 findings	0%

# Reporting

- The computed data quality scores for each data set and data fields, as well as the details on the identified data quality problems can be queried by a set of new Xmeta SQL views.



# Q&A



# Thank You

